Feature Selection in Texts

Magdalena Wiercioch(✉)

Faculty of Mathematics and Computer Science,
Lojasiewicza 6, 30-348 Kraków, Poland
magdalena.wiercioch@ii.uj.edu.pl

Abstract. Feature selection is used in many application areas relevant to expert and intelligent systems, such as machine learning, data mining, cheminformatics and natural language processing. In this study we propose methods for feature selection and features analysis based on Support Vector Machines (SVM) with linear kernels. We explore how these techniques can be used to obtain some interesting information for further exploration of text data. The results provide satisfactory observations which may lead to progress in feature selection field.

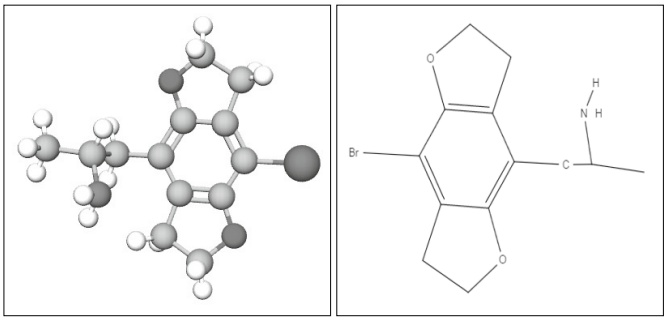
Keywords: Feature selection · Text classification · Dimension reduction · Support Vector Machines

1 Introduction

High dimensional data is a significant problem in both supervised and unsupervised learning [7]. It is becoming even more prominent with the recent explosion of the size of the available datasets both in terms of the number of features in each sample and the number of data samples [17]. The main motivation for reducing the dimensionality of the data and keeping the number of features as low as possible is to decrease the training time and enhance the classification accuracy of the algorithms [3,11].

Take for instance, cheminformatics. Chemical compounds are usually represented by fingerprints, i.e. high dimensional binary strings where a given bit indicates the absence or presence of particular feature of compound [8]. Since a lot of features can be taken into account and it increases the computational costs, another models are considered as well. Chemical databases often store data in form of SMILES. A SMILE (Simplified Molecular Input Line Entry System) is a string of ASCII characters associated with atoms and bonds which build the molecule (see Fig. 1). It is usually visually presented as a graph. Feature selection has received considerable attention in the machine learning and data mining communities. It is typically performed by sorting linguistic features according to some weighting measure [13,16] and then setting up a some kind of threshold on the weights or specifying a percentage or number of highly scored features to be retained. Features with lower weights are omitted as having less significance

M. Kurzynski et al. (eds.), *Proceedings of the 10th International Conference on Computer Recognition Systems CORES 2017*, Advances in Intelligent Systems and Computing 578, DOI 10.1007/978-3-319-59162-9_35



(a) A molecule in 3D space. (b) A graph-like representation of the molecule.

Fig. 1. The SMILES for this molecule is CC(N)Cc1c2CCOc2c(Br)c2CCOc12.

for the classification decision. Furthermore, [14] proposes the method that uses frequent graphs. In addition, much of the past research in cheminformatics has focused on substructure based approaches [10,15]. *The main contribution of this paper is selecting the most valuable fragments of molecules.* However, we also test our approaches on text documents with text stories. In this study we propose techniques for feature selection and analysis based on Support Vector Machines (SVM) [4] with linear kernels. Our experimental results on real world chemical datasets and a collection of Reuters documents [9] show that the demonstrated approaches are promising tools. As it shows, the methods will also be useful for further investigations in the fields connected with representation construction. In the following sections we first describe the basic concepts used in further study. Then, we give a brief explanation of our feature selection methods. Section 3 introduces the experiments and presents the results. We conclude with a brief summary and the outline of future research.

2 Methodology

Our hypothesis in feature selection is that the spatial distribution of text features carries important information regarding the importance of the feature. In presented approaches we consider text data, i.e. SMILES and stories. Figure 2 gives an overview on our approach. In step 1, the text is preprocessed. The type of preprocessing depends on data.

– As previously mentioned, according to SMILES notation, the molecule is represented as a graph. Thus, for each data-molecule random SMILES walks of given length l are performed. In consequence, we obtain many fragments of the graph. Additionally, each fragment has the same label as the graph it belongs to (see Subsect. 3.1). Figure 3 shows two fragments extracted from the original chemical compound. What is more, the preprocessing stage includes a simple tokenization. We assume that SMILES are considered as 2 character

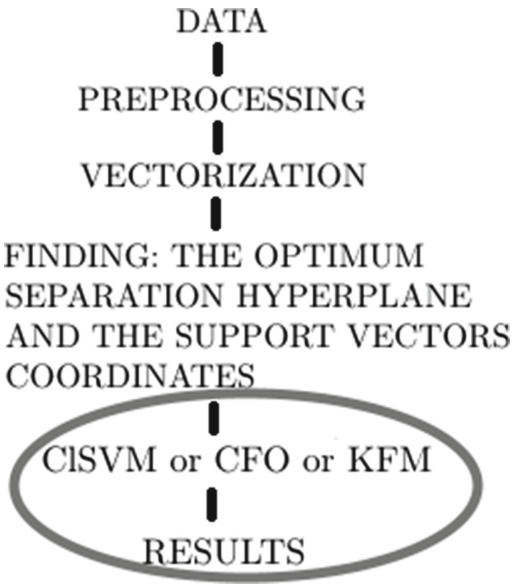


Fig. 2. The procedure workflow.

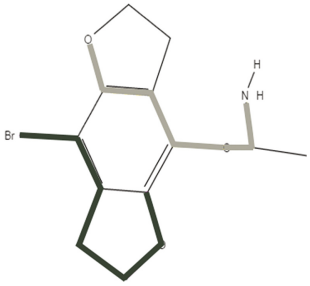


Fig. 3. Visualization of two paths of length 6.

strings. Nevertheless, the chemical properties enforce some exceptions. For instance, [Br-] is treated as a single token.
– As far as texts are considered, they sometimes consist of useful data which has to be removed. These are stopwords or special characters (#, %).

In order to perform some machine learning operations on texts, the conversion of text data into numerical data is required (see the vectorization phase in Fig. 2). Next the LibSVM package [2] is used to train a Support Vector Machine (SVM) [4] for classification. We use the Support Vector Machine with linear kernels. After the vectorization [5], training examples are described by vectors $x_i = (x_i^1, \ldots, x_i^d)$, where d represents the dimensionality of the feature space. The

subsequent step 2 is 1 out of 3 different approaches: Clustered SVM (ClSVM), Counting Features Occurrences (CFO) and Key Features-to-Model (KFM).

Clustered SVM (ClSVM). The classification step provides a set of points, namely support vectors. The randomly selected support vectors are initial cluster centers for k-mean clustering [12]. The goal of such an approach is to gather some features into clusters. It enables to verify the kind of features (in sense of their chemical properties) included together. The whole workflow of ClSVM with the previous steps for chemical data is summarized schematically in Fig. 4.

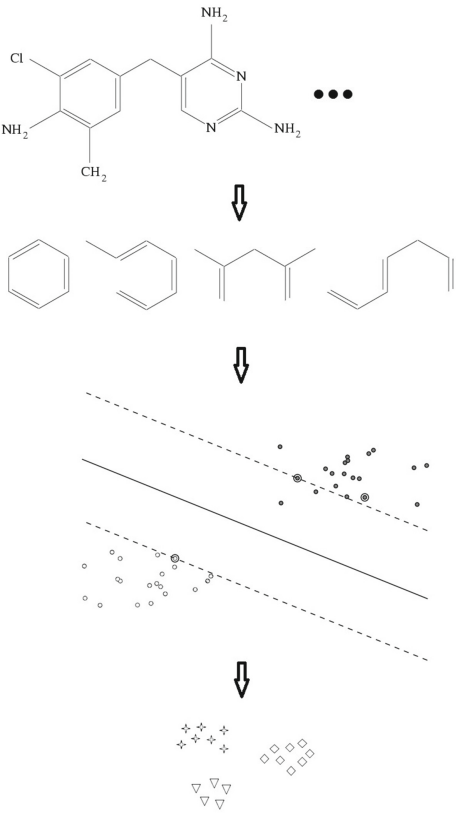


Fig. 4. The procedure workflow with ClSVM.

Key Features-to-Model (KFM). We assume the support vectors provide the most valuable features. Thus, these features are used to create a new representation. To be more precise, for each class we select n most common points from the set of support vectors and concatenate them (see Fig. 5).

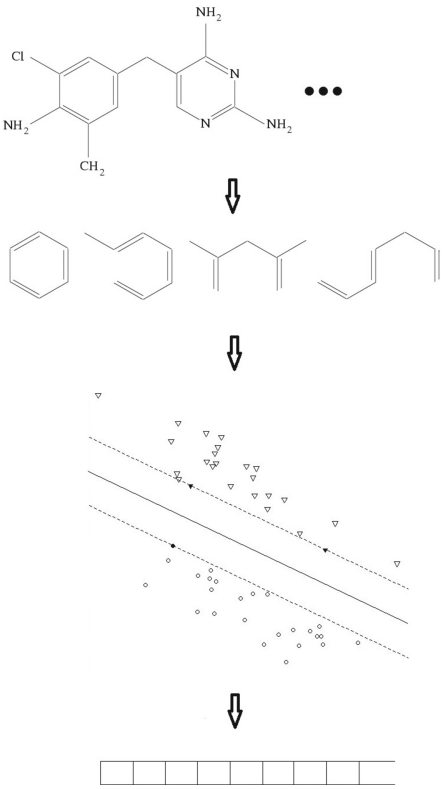


Fig. 5. The procedure workflow with KFM.

Counting Features Occurrences (CFO). Given a linear hyperplane H which divides a dataset X into two regions X_{-1} and X_{+1} we perform a classification of a new point - x_{new} (a new word or a molecule fragment) based on the following system:

- The number of occurrences of the new element within each separate class is calculated: num_{-1}, num_{+1}.
- Finally, the point x_{new} is assigned to the class where it occurs more often.

Figure 6 illustrates this approach.

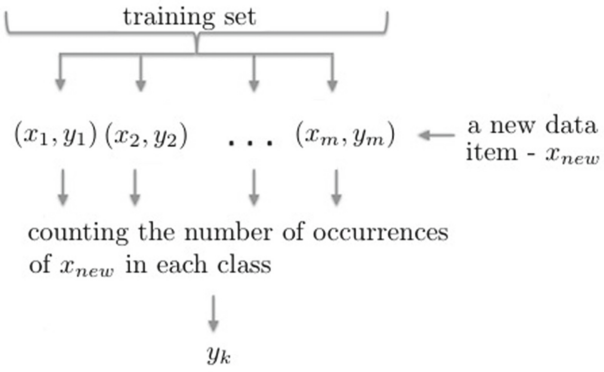


Fig. 6. The procedure workflow with CFO.

3 Experimental Study

In this section we introduce the datasets used in the experiments. Then the presentation of the results is given.

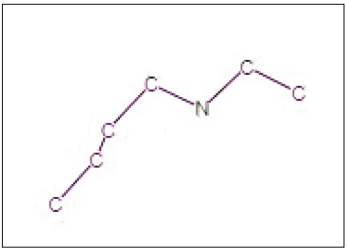
Table 1. Overview of considered data sets. Table contains the names and the number of active and inactive compounds included in initial dataset.

Receptor name	Actives	Inactives
M1	759	938
H1	635	545
$5HT_7$	704	339
$5HT_{2A}$	1835	851
$5HT_6$	1490	341

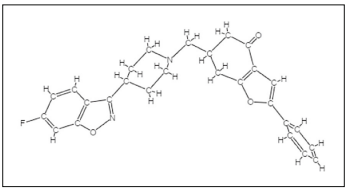
3.1 Data Sets

For experiments we used 2 different types of datasets, i.e. SMILES representation of chemical compounds and text data. As molecules are considered, five biological receptor ligands were used [6], each represents a one receptor ligands, Table 1. Note that there exists an inhibition constant K_i as a kind of activity threshold. For a given molecule, if this factor is less or equal 100nM, the compound is treated as active. However, if K_i is higher than 1000nM, the molecule is seen as inactive. Since the majority of molecules in the real world are inactive, an imbalance dataset problem appears. To tackle it, in our experiments we randomly select inactive molecules as many as actives.

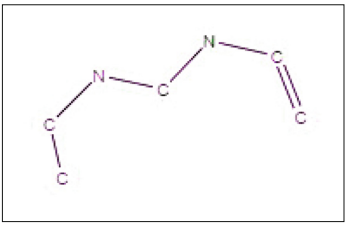
The second collection, namely the Reuters-2000 collection [9] includes a total of 806,791 documents, with news stories covering the period from 20 Aug 1996



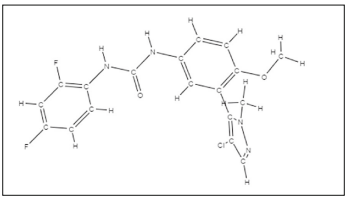
(a) a substructure found in cluster



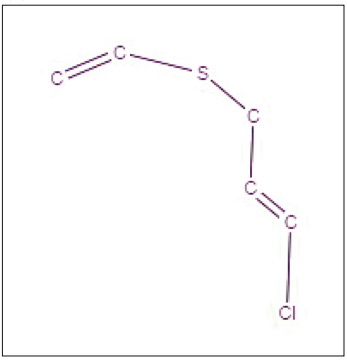
(b) a molecule



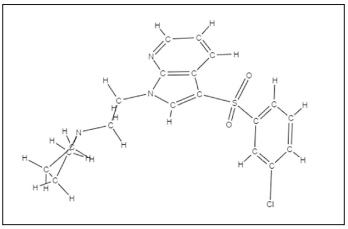
(c) a substructure found in cluster



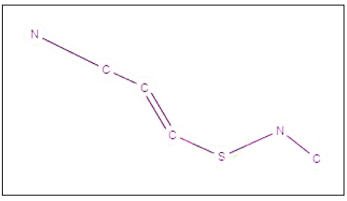
(d) a molecule



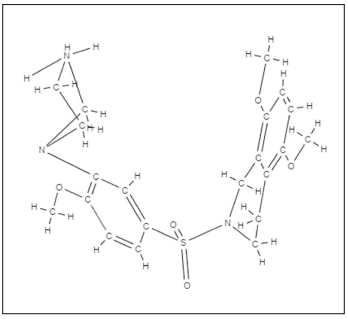
(e) a substructure found in cluster



(f) a molecule



(g) a substructure found in cluster



(h) a molecule

Fig. 7. The examples of substructures found in clusters.

through 19 Aug 1997. We selected documents from these corpus and divided this data into a training and test sets. The finally obtained corpora includes 2 groups of documents on different subjects, i.e. Earn and Acquisition. We modeled a textual document as a kind of a graph of words which corresponds to a graph. Each vertices are associated with unique terms of the document and edges show cooccurrences between the terms within a fixed-size sliding window. The background of this concept is presented in [1].

3.2 Experimental Results

For the tests where chemical data are studied, the obtained graph fragments are the effects of walks of length $l = 6$. The task of conversion of text data into numerical form, namely vectorization, is carried out using CountVectorizer [5]. Moreover, C and other parameters of SVM were tuned using cross validation.

Clustered SVM (ClSVM). The first experiment investigates the results of ClSVM procedure. Indeed, our observations show the obtained clusters include fragments which enable to partially recreate the original molecule. Figure 7 demonstrates only the examples of such fragments found in groups. In is quite interesting since the empirical studies suggest that the fragments included in one cluster enable to reproduce the real compound up to 60% of its real form. Although it is difficult to try to recreate the stories since some data was removed at preprocessing stage, in fact the single clusters contain a set of words associated with the given documents.

Counting Features Occurrences (CFO). Table 2 demonstrates the comparison of CFO classifier with SVM on test data set. The classification quality is calculated with F-measure. As one can see, our approach has performed slightly better in 2 out of 5 cases. It suggests that similar methods are worth to be further explored. For Reuters data we have achieved F-measure of 0.62 which is less than the F-measure of SVM (0.85). This may be caused by type of data.

Key Features-to-Model (KFM). Finally, for KFM we tested the final representation of length 1000. Table 3 shows text classification accuracy for chemical

Table 2. Classification results measured with F-score for CFO and SVM.

Receptor name	CFO	SVM
M1	**0.67**	0.65
H1	0.68	**0.72**
$5HT_7$	0.65	**0.69**
$5HT_{2A}$	0.73	**0.8**
$5HT_6$	**0.73**	0.72

Table 3. Molecules classification accuracy using SVM. The compound are represented with KFM model and typically, as SMILES.

Receptor name	KFM	SVM
M1	0.5	**0.6**
H1	0.6	**0.7**
$5HT_7$	0.55	**0.69**
$5HT_{2A}$	0.7	**0.8**
$5HT_6$	0.68	**0.72**

data using SVM with and without applying the new representation. In fact our model has not achieved better results. However, it still provides a valuable insight into features properties. It should be noticed, the outcomes are not random, so the technique is worth to extend. The experiments outcomes raise an interesting question why it may work. The first reason seems to be connected with the fragments properties - the more fragments, the information is more valuable. What is more, document vectors are sparse and SVM-based approaches are well suited for such problems. Despite the fact our preliminary experiments show encouraging results, this research leaves open a few issues.

- On the exemplary data the methods perform quite satisfactory. However, the more reliable comparison with other similar approaches is absolutely necessary. At this stage, we have not found any analogous techniques.
- Most of the results which are not described in this work are somewhat empirical, i.e. they are based on visual inspection or analysis. Thus, the further exploration should be done.

4 Conclusion

In this paper we have presented three techniques for feature selection and features analysis. The proposed methods seemingly have an influence on text data exploration. In the future, we plan to extend the idea of support vectors-based clusters and analyze another points as potential centroids.

Acknowledgments. This research was partially supported by National Centre of Science (Poland) Grants No. 2016/21/N/ST6/01019.

References

1. Blanco, R., Lioma, C.: Graph-based term weighting for information retrieval. Inf. Retr. **15**(1), 54–92 (2012). http://dx.doi.org/10.1007/s10791-011-9172-x
2. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**, 27:1–27:27 (2011). http://www.csie.ntu.edu.tw/~cjlin/libsvm

3. Chen, R. (ed.): ICICIS 2011, Part II. CCIS, vol. 135. Springer, Heidelberg (2011)
4. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
5. Garreta, R., Moncecchi, G.: Learning Scikit-learn: Machine Learning in Python. Packt Publishing (2013)
6. Gaulton, A., Bellis, L.J., Bento, A.P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., Overington, J.P.: Chembl: a large-scale bioactivity database for drug discovery. Nucleic Acids Res. **40**(D1), D1100 (2011). http://dx.doi.org/10.1093/nar/gkr777
7. Janecek, A., Gansterer, W.N., Demel, M., Ecker, G.: On the relationship between feature selection and classification accuracy. FSDM **4**, 90–105 (2008)
8. Klekota, J., Roth, F.P.: Chemical substructures that enrich for biological activity. Bioinformatics **24**(21), 2518–2525 (2008)
9. Lewis, D.D., Yang, Y., Rose, T.G., Li, F.: Rcv1: a new benchmark collection for text categorization research. J. Mach. Learn. Res. **5**, 361–397 (2004)
10. Kramer, S., De Raedt, L., Helma, C.: Molecular feature mining in HIV data. In: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 136–143 (2001)
11. Lim, T.S., Loh, W.Y., Shih, Y.S.: A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. Mach. Learn. **40**(3), 203–228 (2000)
12. MacQueen, J., et al.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, Oakland, CA, USA, pp. 281–297 (1967)
13. Mladenic, D., Grobelnik, M.: Feature selection for unbalanced class distribution and naive Bayes. In: Proceedings of the 16th International Conference on Machine Learning (ICML), pp. 258–267. Morgan Kaufmann Publishers (1999)
14. Thoma, M., Cheng, H., Gretton, A., Han, J., Kriegel, H.P., Smola, A., Song, L., Yu, P., Yan, X., Borgwardt, K.: Near-optimal supervised feature selection among frequent subgraphs, pp. 1076–1087. Max-Planck-Gesellschaft/Society for Industrial and Applied Mathematics, Philadelphia, May 2009
15. Wale, N., Watson, I.A., Karypis, G.: Comparison of descriptor spaces for chemical compound retrieval and classification. Knowl. Inf. Syst. **14**(3), 347–375 (2008)
16. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: ICML 1997, pp. 412–420 (1997)
17. Zhang, Y., Yang, C., Yang, A., Xiong, C., Zhou, X., Zhang, Z.: Feature selection for classification with class-separability strategy and data envelopment analysis. Neurocomputing 166, 172–184 (2015), http://www.sciencedirect.com/science/article/pii/S0925231215004609